

A short review and primer on the use of human voice in human computer interaction applications

Kristian Lukander

Quantified Employee unit, Finnish Institute of Occupational Health,
 kristian.lukander@ttl.fi,
 POBox 40, Helsinki, 00250, Finland

Abstract. The application of psychophysiology in human-computer interaction is a growing field with significant potential for future smart personalised systems. Working in this emerging field requires comprehension of an array of physiological signals and analysis techniques. Human speech affords, alongside linguistic content, rich information in the intonation, voice quality, prosody, and rhythmic variation of utterances, allowing listeners to recognise numerous distinct emotional states in the speaker. Several types of factors affect speech, ranging from emotions to cognitive load and pathological conditions, providing a promising non-intrusive source for online understanding of context and psychophysiological state.

This paper aims to serve as a primer for the novice, enabling rapid familiarisation with the latest core concepts. We put special emphasis on everyday human-computer interface applications to distinguish from the more common clinical or sports uses of psychophysiology.

This paper is an extract from a comprehensive review of the entire field of ambulatory psychophysiology, including 12 similar chapters, plus application guidelines and systematic review. Thus any citation should be made using the following reference:

B. Cowley, M. Filetti, K. Lukander, J. Tornainen, A. Henelius, L. Ahonen, O. Barral, I. Kosunen, T. Valtonen, M. Huotilainen, N. Ravaja, G. Jacucci. *The Psychophysiology Primer: a guide to methods and a broad review with a focus on human-computer interaction*. Foundations and Trends in Human-Computer Interaction, vol. 9, no. 3-4, pp. 150–307, 2016.

Keywords: eye tracking, psychophysiology, human-computer interaction, primer, review

Human voice production originates at the larynx, where air pressure from the lungs causes vibration of the vocal folds, thereby generating a complex but patterned sound source composed of a fundamental frequency and multiple harmonics. This signal is then filtered through the vocal tract airways (oral and nasal cavities). This vocal apparatus produces a complex interactive system capable of generating a wide variety of sounds [Ghazanfar and Rendall, 2008].

Humans have evolved a nonverbal communication system in which, alongside linguistic content, speech carries rich information in the intonation, voice quality, prosody, and rhythmic variation of utterances, allowing listeners to recognise numerous distinct emotional states in the speaker. Several types of factors, from emotions to cognitive load and pathological conditions, affect the functioning of the larynx, or ‘voice box’, whereby the internal state of the speaker causes tightening/relaxation of the vocal folds, which modulates the acoustic and rhythmic components of speech.

Now that computing systems are starting to listen actively to people (as with Google’s ‘Voice’, Microsoft’s ‘Cortana’, and Apple’s ‘Siri’), human speech provides a promising source for online understanding of context and psychophysiological state with measurement that remains minimally intrusive.

1 Background

Evaluating stress, affect, and mood on the basis of the human voice is not a new notion. Much of the literature on the effects of psychophysiological states on the acoustics of speech production has its roots in the 1970s, in work on stress and lie detectors in interrogation, military, and aerospace settings. Technologies used in these settings – namely, voice stress analysis, voice risk analysis, and layered voice analysis – use the recorded features, including ‘micro tremors’, in a person’s voice to construct a scorable ‘voice gram’, which is then evaluated by a specialist. While these approaches have been utilised in courtrooms and operator monitoring in demanding work tasks, their reliability has been disputed – for these methods, the most convincing results have been obtained in conditions of extreme stress, such as under threat of injury or great operational risk. The recording environments in such research have possessed heterogeneous acoustic characteristics, and the results and metrics from the relevant studies cannot be cross-evaluated reliably [Hopkins et al., 2005, Harnsberger et al., 2009].

However, even everyday user interaction situations appear to elicit strong enough emotional and stress responses to produce systematic, detectable changes in voice parameters. In the interactive setting of HCI specifically, the psychophysiological states recognised in the literature as having an effect on speech production are cognitive workload, or ‘stress’ [Lively et al., 1993]; physical stress [Godin and Hansen, 2015], and various emotional states [El Ayadi et al., 2011].

2 Methods

Changes in human voice production can be measured with a microphone, in combination with the use of mathematical models to associate acoustic changes with the functioning of the larynx, or through electroglottography, in which the system uses two pairs of electrodes (one pair on either side of the subject’s throat) to measure the variations over time in the degree of contact of the vocal folds during voice production [Kania et al., 2006]. When microphones are employed, the approach typically involves inverse-filtering microphone recordings to model

the waveform of the glottal airflow pulses, which, in turn, reflect the movements of the vocal folds [Alku, 2011]. There is a downside, however: microphone recordings vary greatly in quality and with respect to noise parameters related to microphone type, environmental factors, and distance from the speaker. Guidelines for selection of microphones suitable for human voice research are provided in a summary by Svec and Granqvist [2010].

Several derived low-level features and combinations thereof have been suggested to be correlated with variations in internal states. Table 1, compiled on the basis of work by Ververidis and Kotropoulos [2006] and Scherer [2003], presents the effects of fundamental emotional state on selected features. Typical features derived from speech signals include loudness, fundamental frequency, word and utterance rate (speed), jitter, zero-crossing rate, and frequency ratios. While acoustic features of speech such as pitch, timing, voice quality, and articulation have been shown to correlate highly with underlying emotional activation (low–high), there is no agreement as to how these dimensions correlate with the valence of the emotion [El Ayadi et al., 2011]. Instead of direct comparisons involving individual features or combinations of them, modern approaches tend to use machine learning methods to improve detection rates [Zhou et al., 2001].

Table 1. Effects of certain fundamental emotional states on selected acoustic features, compiled from work by Scherer [2003] and Ververidis and Kotropoulos [2006].

		Stress	Anger/rage	Fear/panic	Disgust	Sadness	Joy/elation	Boredom
Intensity	Mean	↗	↗	↗	↘	↘	↗	
	Range		↗	↗		↘	↗	
F0 ¹	floor/mean	↗	↗	↗	↘	↘	↗	
	variability		↗			↘	↗	↘
	range		↗	↗	↗ ²	↘ ³	↗	↘
Sentence contours			↘	↗		↘	↘	
High-frequency energy			↗	↗		↘	↗	
Speech and articulation rate			↗ ⁴	↗	↘	↗ ⁵	↗	↘
Duration			↘	↘		↗	↘	

¹ Fundamental frequency

² According to Ververidis and Kotropoulos [2006], the F0 range among males increases.

³ Studies do not agree.

⁴ According to Scherer [2003], the articulation rate among males decreases.

⁵ According to Ververidis and Kotropoulos [2006], the articulation rate among males increases.

3 Applications

While variations between one speaker and the next require calibration and baselines, online calculation of acoustic parameters and spectral measures is relatively easy and robust. In one recent innovation, a set of open-source toolboxes for automated feature extraction and voice analysis has been developed. These include openSMILE – the Munich Versatile and Fast Open-Source Audio Feature Extractor [Eyben et al., 2010] – and AMMON (Affective and Mental Health Monitor) [Chang et al., 2011].

A considerable amount of the recent research in this field has been presented in connection with the Interspeech computational paralinguistic challenges [Schuller et al., 2015]. Since 2009, these challenges have called for methods of evaluating speech for the detection of age, gender, affect and emotional parameters, personality, likeability, pathologies and diseases, social signals, signs of conflict, cognitive load, and physical demands. For an extensive list of examples of computational solutions for the paralinguistic detection of cognitive and emotional states, the reader is directed to the Interspeech repository available online at <http://compare.openaudio.eu/>.

4 Conclusion

For the interactive setting that is the focus of our attention, analysing the acoustics of speech production offers a non-intrusive online metric for gauging the internal state of the user. There is considerable potential on account of the pervasiveness and unobtrusive nature of the method. The classification performance of automated solutions is beginning to reach an acceptable level of sensitivity and reliability, at least upon user-specific calibration to accommodate the effects of differences in languages and dialects, individual-to-individual differences in speech production, and variations in stress and affective responses.

Another source of motivation for significant improvements in both recognising and producing natural paralinguistic cues associated with empathetic responses arises from future needs related to affective computing, robotics, and artificial conversation partners in general. These technologies necessitate naturalistic input and output down to the smallest detail in order to escape the ‘uncanny valley’ [Mori et al., 2012].

Bibliography

- Paavo Alku. Glottal inverse filtering analysis of human voice production—a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36(5):623–650, 2011.
- Keng-hao Chang, Drew Fisher, John Canny, and Björn Hartmann. How’s my mood and stress?: An efficient speech analysis library for unobtrusive monitoring on mobile phones. In *Proceedings of the 6th International Conference on Body Area Networks*, pages 71–77. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011.
- Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.
- Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.
- Asif A Ghazanfar and Drew Rendall. Evolution of human vocal production. *Current Biology*, 18(11):R457–R460, 2008.
- Keith W Godin and John HL Hansen. Physical task stress and speaker variability in voice quality. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):1–13, 2015.
- James D Harnsberger, Harry Hollien, Camilo A Martin, and Kevin A Hollien. Stress and deception in speech: Evaluating layered voice analysis*. *Journal of forensic sciences*, 54(3):642–650, 2009.
- Clifford S Hopkins, Roy J Ratley, Daniel S Benincasa, and John J Grieco. Evaluation of voice stress analysis technology. In *System Sciences, 2005. HICSS’05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 20b–20b. IEEE, 2005.
- Romain E Kania, Dana M Hartl, Stéphane Hans, Shinji Maeda, Jacqueline Vaissiere, and Daniel F Brasnu. Fundamental frequency histograms measured by electroglottography during speech: A pilot study for standardization. *Journal of voice*, 20(1):18–24, 2006.
- Scott E Lively, David B Pisoni, W Van Summers, and Robert H Bernacki. Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences. *The Journal of the Acoustical Society of America*, 93(5):2962–2973, 1993.
- Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *Robotics & Automation Magazine, IEEE*, 19(2):98–100, 2012.
- Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1):227–256, 2003.
- B Schuller, S Steidl, and Batliner A. Repository of the inter-speech computational paralinguistics challenge (compare) series. <http://www.compare.openaudio.eu/>, 2015. Accessed:2016-01-15.

- Jan G Svec and Svante Granqvist. Guidelines for selecting microphones for human voice production research. *American Journal of Speech-Language Pathology*, 19(4):356–368, 2010.
- Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181, 2006.
- Guojun Zhou, John HL Hansen, and James F Kaiser. Nonlinear feature based classification of speech under stress. *Speech and Audio Processing, IEEE Transactions on*, 9(3):201–216, 2001.